Journal of Multivariate Analysis 112 (2012) 48–62

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Uniqueness of linear factorizations into independent subspaces

Harold W. Gutch [a,c,*], Fabian J. Theis [b,c]

[a] Max Planck Institute for Dynamics and Self-Organization (MPIDS), 37077 Goettingen, Germany

[b] CMB, Institute of Bioinformatics and Systems Biology, German Research Center for Environmental Health, 85764 Neuherberg, Germany

[c] Technical University of Munich, Arcisstrasse 21, 80333 München, Germany

ARTICLE INFO

Article history:
Received 26 April 2010
Available online 19 June 2012

AMS 2000 subject classifications:
62E10
62H25

Keywords:
Statistical independence
Independent component analysis
Independent subspace analysis
Separability
Inverse models

ABSTRACT

Given a random vector $\mathbf{X}$, we address the question of linear separability of $\mathbf{X}$, that is, the task of finding a linear operator $\mathbf{W}$ such that we have $(\mathbf{S}_1, \ldots, \mathbf{S}_M) = (\mathbf{WX})$ with statistically independent random vectors $\mathbf{S}_i$. As this requirement alone is already fulfilled trivially by $\mathbf{X}$ being independent of the empty rest, we require that the components be not further decomposable. We show that if $\mathbf{X}$ has finite covariance, such a representation is unique up to trivial indeterminacies. We propose an algorithm based on this proof and demonstrate its applicability. Related algorithms, however with fixed dimensionality of the subspaces, have already been successfully employed in biomedical applications, such as separation of fMRI recorded data. Based on the presented uniqueness result, it is now clear that also subspace dimensions can be determined in a unique and therefore meaningful fashion, which shows the advantages of independent subspace analysis in contrast to methods like principal component analysis.

1. Introduction

Assume a random vector $\mathbf{S}$ consisting of statistically independent components S_i, none of which is Gaussian (normally distributed). If the components of a linear mixing $\mathbf{AS}$ then again are statistically independent, one can show that $\mathbf{A}$ is at most the product of a permutation and scaling within the components, which originally was shown using the Darmois–Skitovitch theorem [11,21,30]. Under the additional assumption of finite covariance of $\mathbf{S}$, one may additionally allow at most one of the S_i to be Gaussian [10]. The assumption of finite covariance actually is not required for this to hold [13], but if one assumes it, a simpler proof is possible, based on the idea that the characteristic function of $\mathbf{S}$ factorizes, so its logarithm has a diagonal Hessian almost everywhere [33].

This property of random variables has driven the development of algorithms performing the so-called Independent Component Analysis (ICA) under some approximations of statistical independence [4,10,20,37]. Exact cost functions, the so-called contrasts such as mutual information, are difficult to estimate in practice, where the random vector in question is only known up to some finite precision, so many approximations have been extensively studied. Such algorithms have been successfully used in various fields, e.g. signal processing, biomedical imaging and analysis of financial data, where it was argued that the data sets to be analyzed can be approximated well enough by modeling them as random variables mixed in a linear fashion; see [9,18] and references therein.

Apart from the question of validity when transferring the mathematical theory to real life data sets, another problem is apparent: What if a given random vector $\mathbf{X}$ has no such representation? This motivates the question if the original claim has a straight-forward extension to higher dimensions: if we write $(\mathbf{S}_1, \ldots, \mathbf{S}_M) := \mathbf{WX}$ with independent random vectors

* Corresponding author at: Max Planck Institute for Dynamics and Self-Organization (MPIDS), 37077 Goettingen, Germany.
E-mail addresses: harold.gutch@ds.mpg.de (H.W. Gutch), fabian.theis@helmholtz-muenchen.de (F.J. Theis).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.05.019

$\mathbf{S}_i$ of which at most one is Gaussian, again, are these unique up to permutation and invertible linear transformations (the multidimensional translation of scaling) within the $\mathbf{S}_i$? Loosened assumptions to the ICA model where specifically such subspaces in data were sought for have gained a lot of interest in recent years; see e.g. [3,8,14,24,28,40]. The task of finding a basis in which a random vector $\mathbf{X}$ has this property is usually denoted Independent Subspace Analysis (ISA), as one typically reads the independent random vectors gained here as data subsets or data subspaces [7]. Obviously this task requires some minimality constraint, as, given such an independent representation, we might arbitrarily group together some of the components and thus of course get two representations differing in more than just permutation and linear transformations. Our minimality constraint is the inability to decompose any of the components even further, a property we call *irreducibility* of the components.

Our main result is the following uniqueness theorem.

Theorem 1.1. *The decomposition of a random vector* $\mathbf{X}$ *with existing covariance into independent, irreducible components is unique up to order and invertible transformations within the components and an invertible transformation in the possibly higher-dimensional Gaussian component.*

Transferring this mathematical statement to the real world, the conclusion now is that *every* kind of data that one can model as a random vector inherently has a unique factorization into subspaces. This statement has far wider applicability than ICA, which states uniqueness of a factorization only *if* one exists at all within the limits of the model—i.e. if there is a decomposition into independent one-dimensional components. In contrast to this, ISA is applicable to almost any kind of high-dimensional data that can be modeled as a random vector—the only requirement is finite covariance.

This manuscript is structured as follows. In Section 2 we define the notation used and the framework we work in and state a few simple lemmata used. The main part of this manuscript, Section 3, consists of the proof of Theorem 1.1. In Section 4, we propose an algorithm whose main ideas are based on the theoretical proof and demonstrate its applicability in a simulation study. Finally, in Section 5 we shed some light on the practical usefulness of this result, compare it with the literature and address some open questions.

Some parts of this work were presented at the ICA 2007 conference [15], stating Theorem 1.1, however lacking the proof and the algorithmic approach.

2. Definition of independent subspace analysis

We will define the analyzed model and review a few properties of characteristic functions. We restrict this analysis to the real case, that is, real valued random vectors and real linear mixings thereof, although extensions to the complex case are possible.

2.1. Notation

In order to be able to quickly differentiate between scalars and vectors, scalars are depicted in regular font, e.g. $x \in \mathbb{R}$ while vectors and matrices are depicted in bold font, e.g. $\mathbf{x} \in \mathbb{R}^n$. Random values and vectors are always depicted in uppercase letters, e.g. S and $\mathbf{S}$, and we will only need the two letters $\mathbf{S}$, and $\mathbf{X}$ (and regular typeface versions thereof) for these; all other uppercase letters used represent real valued matrices. In order to keep the notation as simple as possible, vectors will often be written in rows, e.g. $\mathbf{x} = \mathbf{A}(\mathbf{v}_1, \mathbf{v}_2)$ instead of $\mathbf{x} = \mathbf{A}(\mathbf{v}_1^\top, \mathbf{v}_2^\top)^\top$. The symbol ∂_i denotes the i-th partial derivative operator, so for a function depending on $\mathbf{x} = (x_1, \ldots, x_n)$, we have $\partial_i = \frac{\partial}{\partial x_i}$. We write $\mathbf{d}f$ for the differential of an $f \in C^1$, and we depict the Hessian of an $f \in C^2$ with $\mathbf{H}_f$, that is $\mathbf{e}_i^\top \mathbf{H}_f \mathbf{e}_j = \partial_i \partial_j f$. We write $\mathbf{d}f|_{\mathbf{x}}$ instead of $(\mathbf{d}f)(\mathbf{x})$, the differential of f evaluated at $\mathbf{x}$, and similarly $\mathbf{H}_f|_{\mathbf{x}}$ instead of $\mathbf{H}_f(\mathbf{x})$, the Hessian of f evaluated at $\mathbf{x}$.

2.2. Irreducibility

Let us now introduce the key notion of irreducibility and point out the special role of Gaussian random vectors.

Definition 2.1. An n-dimensional random vector $\mathbf{X}$ is said to be *reducible* if it can be written as $\mathbf{X} = \mathbf{A}(\mathbf{S}_1, \mathbf{S}_2)$ with some invertible $n \times n$-matrix $\mathbf{A}$, a k-dimensional random vector $\mathbf{S}_1$ and an $(n-k)$-dimensional random vector $\mathbf{S}_2$, where $\mathbf{S}_1$ is independent of $\mathbf{S}_2$. A random vector that is not reducible is called *irreducible*.

Remark 2.1. For example, any n-dimensional Gaussian random vector is reducible if $n > 1$: Gaussians are fully defined by their first and second order moments, so here independence is equivalent to decorrelation, and for every random vector $\mathbf{X}$ with finite covariance there is some invertible matrix $\mathbf{A}$ such that $\mathbf{AX}$ is decorrelated (see Lemma 3.1). Therefore, an n-dimensional Gaussian can always be fully reduced to one-dimensional components.

Obviously both properties, irreducibility and reducibility, are preserved under any invertible linear transformation.

A decomposition $(\mathbf{X}_1, \ldots, \mathbf{X}_L) = \mathbf{X}$ of a random vector $\mathbf{X}$ is said to be *independent*, if the random vectors $\mathbf{X}_j$ $(j = 1, \ldots, L)$ are mutually statistically independent. It is said to be *irreducible*, if the vectors $\mathbf{X}_j$ additionally are irreducible.

Remark 2.2. It is straight-forward to see that for any random vector $\mathbf{X}$ there is some invertible matrix $\mathbf{A}$ such that $\mathbf{AX} = (\mathbf{X}_1, \ldots, \mathbf{X}_L)$ is an irreducible decomposition: either $\mathbf{X}$ already is irreducible or there is some invertible $\mathbf{A}$ such that $\mathbf{AX} = (\mathbf{X}_1, \mathbf{X}_2)$ with independent $\mathbf{X}_1, \mathbf{X}_2$. If these two are irreducible, we are finished, otherwise we proceed to decompose whichever of the two still is reducible. After a finite number $L < \dim(\mathbf{X})$ of steps, we are left with irreducible components.

Having established the existence of such a decomposition, we define a normalized version of it.

Definition 2.2. Assume an n-dimensional random vector $\mathbf{X}$ and an invertible $n \times n$ matrix $\mathbf{A}$ such that $\mathbf{S} = \mathbf{AX}$ can be subdivided into $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_M)$ where

(i) $\dim(\mathbf{S}_i) \leq \dim(\mathbf{S}_j)$ for any $1 \leq i < j \leq M$,
(ii) the random vectors $\mathbf{S}_k$ are mutually independent,
(iii) at most one of the $\mathbf{S}_k$ is Gaussian,
(iv) all non-Gaussian $\mathbf{S}_k$ are irreducible,

then, with $\mathbf{m} := (\dim(\mathbf{S}_1), \ldots, \dim(\mathbf{S}_m))$, an ordered partition of $\dim(\mathbf{S})$, the pair $(\mathbf{A}, \mathbf{m})$ is called an Irreducible Subspace Analysis (ISA) of $\mathbf{X}$ and the random vectors $\mathbf{S}_i$ are called the irreducible components of $(\mathbf{A}, \mathbf{m})$.

Note that we have gathered all independent one-dimensional Gaussians into a single, higher-dimensional Gaussian component, an idea that was introduced in [5,6]. Here, observe that any two-dimensional rotation maps two independent Gaussians again onto two independent Gaussians, a fact that also holds for higher dimensions. Therefore, it is always possible for two ISAs of a random vector $\mathbf{X}$ to differ in the matrix component by a rotation in the higher-dimensional Gaussian, so the irreducible components of the Gaussian are also unique only up to this indeterminacy.

We note that according merely to the definition, a given $\mathbf{X}$ may have several ISAs, differing in either the basis $\mathbf{A}$, the sizes $\mathbf{m}$ or both.

2.3. The characteristic function and its properties

In the following we will work extensively with the characteristic function of a random vector, so we shortly review its definition and some elementary properties.

Definition 2.3. Let $\mathbf{X}$ be an n-dimensional random vector. Then the *characteristic function* of $\mathbf{X}$ is defined as $\widehat{\mathbf{X}}(\mathbf{x}) := E\{\exp(i\mathbf{X}^\top \mathbf{x})\}$ where $\mathbf{x} \in \mathbb{R}^n$.

The characteristic function has similar properties to the density when it comes to statistic independence: assume $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent. Then their joint characteristic function is equal to the product of the marginal characteristic functions [16]:

$$\widehat{\mathbf{X}}(\mathbf{x}) = \widehat{\mathbf{X}}_1(\mathbf{x}_1)\widehat{\mathbf{X}}_2(\mathbf{x}_2).$$

A local logarithm (note that $\widehat{\mathbf{X}}$ is complex valued) of $\widehat{\mathbf{X}}$ – this logarithm is also known as the second characteristic function [39] – then splits into the sum of the local logarithms of the marginal characteristic functions. Characteristic functions always exist, whereas not all random variables admit a density. If $\mathbf{A}$ is an invertible matrix and $\mathbf{S}$ is a random vector, then

$$\widehat{\mathbf{AS}}(\mathbf{x}) = E\left(\exp(i\mathbf{S}^\top \mathbf{A}^\top \mathbf{x})\right) = \widehat{\mathbf{S}}(\mathbf{A}^\top \mathbf{x}).$$

The characteristic function of a random vector has a simple connection to its moments (given that they exist). If $\mathbf{X}$ is a random vector then its $(n_1, \ldots, n_k)$-th moment can be calculated as follows (see e.g. [16]):

$$E\{X_1^{n_1} \cdots X_k^{n_k}\} = i^{-m} \frac{\partial^m}{\partial x_1^{n_1} \cdots \partial x_k^{n_k}} \widehat{\mathbf{X}}(\mathbf{x})|_0 \tag{1}$$

with $m = n_1 + \cdots + n_k$.

Among all random variables, the ones with the simplest representation are Gaussians. For example, a one dimensional random variable X with known density $p_X > 0$ is Gaussian if and only if $\ln p_X$ is a polynomial of degree at most 2, i.e. if $(\ln p_X)'' = 0$. This property can be used to extract Gaussian components from larger random vectors [38].

2.4. Complex differentiation and linear transformations

In the main proof, we will iteratively extract components. For this we make use of the following three lemmata, the proofs of which are omitted as they are straight-forward.

Lemma 2.1 (*Differential of a Product*). *Assume twice differentiable functions $f_k : \mathbb{R}^n \to \mathbb{C}$ $(k = 1, \ldots, M)$ and (not necessarily different) differential operators ∂_i and ∂_j. Then*

$$f\partial_i\partial_j f - (\partial_i f)(\partial_j f) = \sum_{k=1}^{M} \left(\prod_{l \neq k} f_l^2 \right) [f_k \partial_i \partial_j f_k - (\partial_i f_k)(\partial_j f_k)]$$

where $f : \mathbb{R}^n \to \mathbb{C}$ is defined by $f(\mathbf{x}) := \prod_{k=1}^{M} f_k(\mathbf{x})$.

Lemma 2.2 (*Directional Derivative and Hessian*)**.** *Let $g(\mathbf{x}) := f(\mathbf{A}(\mathbf{x}))$ where $\mathbf{A}$ is an $(m \times n)$-matrix over $\mathbb{R}$ and f is a twice differentiable function on $\mathbb{R}^m$. Then for every $\mathbf{x} \in \mathbb{R}^n$*

(i) $(\partial_i g)(\mathbf{x}) = \mathbf{d}f|_{\mathbf{Ax}}\mathbf{a}_i$
(ii) $(\partial_i\partial_j g)(\mathbf{x}) = (\mathbf{a}_i)^\top \mathbf{H}_f|_{\mathbf{Ax}}\mathbf{a}_j$

where $\mathbf{a}_i$ denotes the i-th column of $\mathbf{A}$.

Lemma 2.3 (*Logarithmic Hessian*)**.** *Assume $U \subset \mathbb{R}^n$ and $f : U \to \mathbb{C}$ to be a twice continuously differentiable function with some $\mathbf{x} \in U$ such that $f(\mathbf{x}) \neq 0$. Then, in some neighborhood of $\mathbf{x}$,*

$$\mathbf{H}_f = f\mathbf{H}_g + f(\mathbf{d}g)^\top(\mathbf{d}g)$$

where $g := \log f$ is a local complex logarithm.

3. Uniqueness of ISA

We will now show Theorem 1.1 in a number of steps. Without loss of generality we assume $\mathbf{X}$ to be centered (zero-mean). As the existence of an ISA of $\mathbf{X}$ holds, we may additionally assume to already have one such decomposition.

Assumption A1. Let $\mathbf{X}$ be an n-dimensional random vector with mean 0 and finite covariance and $(\mathbf{A}, \mathbf{m})$ be an ISA of $\mathbf{X}$ with irreducible components $\mathbf{S}_1, \ldots, \mathbf{S}_M$.

The claim of Theorem 1.1 is equivalent to the claim that for any other ISA $(\mathbf{A}', \mathbf{m}')$ of $\mathbf{X}$, actually $\mathbf{m}' = \mathbf{m}$ and $(\mathbf{A}'\mathbf{A}^{-1})$ can be written as the product of an invertible block-diagonal matrix with blocks of size $m_1, \ldots, m_M$ and a block-permutation matrix, swapping at most blocks of the same size corresponding to non-Gaussians. In terms of the irreducible components of $(\mathbf{A}, \mathbf{m})$ and $(\mathbf{A}', \mathbf{m}')$, this is equivalent to them being mapped to each other by such a product.

3.1. The Gaussian subspace

Recently, the idea of Non-Gaussian Component Analysis (NGCA), or Non-Gaussian Subspace Analysis (NGSA), has been proposed, where one separates a higher dimensional distribution into two independent parts, one of them being a high dimensional Gaussian (the *Gaussian subspace*) and the rest (the *Non-Gaussian subspace*) [5,6]. If the Gaussian subspace is maximal (i.e. there is no way to split off an independent Gaussian from the Non-Gaussian subspace), this decomposition is unique up to transformations within each of the two subspaces.

In order to simplify notation, we from now on will use the following convention.

Definition 3.1. Two random vectors $\mathbf{X}$ and $\mathbf{S}$ are called *equivalent* (in symbols: $\mathbf{X} \sim \mathbf{S}$) if there is some invertible $\mathbf{A}$ such that $\mathbf{X} = \mathbf{AS}$.

The following theorem establishes uniqueness of the decomposition.

Theorem 3.1 (*Uniqueness of NGSA*)**.** *Assume $\mathbf{X}$, a random vector with existing covariance and two arbitrary decompositions $\mathbf{X} = \mathbf{A}(\mathbf{X}_N, \mathbf{X}_G) = \mathbf{B}(\mathbf{S}_N, \mathbf{S}_G)$ such that*

(i) *$\mathbf{X}_G$ and $\mathbf{S}_G$ are higher-dimensional Gaussians,*
(ii) *$\mathbf{X}_N$ and $\mathbf{X}_G$ are independent and so are $\mathbf{S}_N$ and $\mathbf{S}_G$,*
(iii) *the decompositions are maximally reduced, in the sense that there is no projection $\mathbf{M}$ such that the first component of $\mathbf{MX}_N$ is a Gaussian independent of the rest, and similarly for $\mathbf{S}_N$.*

Then $\mathbf{X}_N \sim \mathbf{S}_N$ and $\mathbf{X}_G \sim \mathbf{S}_G$.

For a proof of this theorem, we refer the reader to [38]. We note that as a special case this also includes deterministic components, which can be seen as Gaussians with variance 0.

This theorem shows that both the maximally Gaussian subspace and the rest are essentially unique. It therefore suffices to show uniqueness of ISA for the non-Gaussian component of $\mathbf{X}$, that is, the part of $\mathbf{X}$ that contains no independent Gaussian, and we may restrict our analysis to random vectors of this kind. It will furthermore be of use in the following to assume $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ with independent $\mathbf{X}_1, \mathbf{X}_2$.

Assumption A2. Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ such that $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent and such that for no $\mathbf{B} \in \mathrm{Gl}(m)$ the projection $(\mathbf{BX})_1$ is normal and independent of the rest of $(\mathbf{BX})$.

If both A1 and A2 hold, then obviously none of the components $\mathbf{S}_i$ are normally distributed.

3.2. Whitening

We will now show that we may assume $\mathbf{X}$ and $\mathbf{S}$ to be decorrelated, which implies $\mathbf{A}$ being orthonormal. The proofs in this section hold in a more general setting than covered by our current assumptions A1 and A2.

Lemma 3.1. *Assume an n-dimensional random vector* $\mathbf{X}$*. Then there is an invertible* $(n \times n)$ *matrix* $\mathbf{T}$ *such that* $\mathrm{Cov}(\mathbf{TX}) = \mathbf{I}$.

Whitening is a simple application of the eigenvector-decomposition of the (positive-semidefinite) covariance; see e.g. [18, Section 6.4].

Let us now bring this to use.

Lemma 3.2. *Assume* $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_M)$ *with existing covariance and independent random vectors* $\mathbf{S}_k$*, such that for no invertible* $\mathbf{B}$*, the projection* $(\mathbf{BS})_1$ *is deterministic. Assume furthermore an invertible* $\mathbf{A}$ *and* $\mathbf{X} := \mathbf{AS}$ *where* $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_L)$ *with independent* $\mathbf{X}_j$*. Then there are* $\mathbf{S}'_k \sim \mathbf{S}_k$ $(k = 1, \ldots, M)$*,* $\mathbf{X}'_j \sim \mathbf{X}_j$ $(j = 1, \ldots, L)$*, and an orthonormal* $\mathbf{A}'$ *such that every* $\mathbf{S}'_k$ *and* $\mathbf{X}'_j$ *is decorrelated and* $(\mathbf{X}'_1, \ldots, \mathbf{X}'_L) = \mathbf{A}'(\mathbf{S}'_1, \ldots, \mathbf{S}'_M)$.

Proof. As $\mathbf{S}$ has existing covariance, so does $\mathbf{X} = \mathbf{AS}$. Using (3.1), we may first decorrelate every component of $\mathbf{X}$ and $\mathbf{S}$, modifying $\mathbf{A}$ accordingly. We here scale our decorrelation matrices such that after decorrelation we have unit covariance everywhere (this can be done due to the assumption of non-deterministic projections). As this operation was performed fully within the single components, for the decorrelated $\mathbf{S}'_k$ and $\mathbf{X}'_j$, we have $\mathbf{S}'_k \sim \mathbf{S}_k$ and $\mathbf{X}'_j \sim \mathbf{X}_j$. Then, setting $\mathbf{S}' := (\mathbf{S}'_1, \ldots, \mathbf{S}'_M)$, $\mathbf{X}' := (\mathbf{X}'_1, \ldots, \mathbf{X}'_L)$, and letting $\mathbf{A}'$ be the modified $\mathbf{A}$, we have

$$\mathbf{I} = \mathrm{Cov}(\mathbf{X}') = \mathrm{Cov}(\mathbf{A}'\mathbf{S}') = \mathbf{A}'\mathrm{Cov}(\mathbf{S}')\mathbf{A}'^{\top} = \mathbf{A}'\mathbf{I}\mathbf{A}'^{\top}$$

so $\mathbf{A}'$ is orthonormal. □

Given our assumptions A1 and A2, decorrelating the independent components $\mathbf{X}_j$ and $\mathbf{S}_k$ this way does not lose any generality: the independence of the components still holds, and performing decorrelation in this manner we may therefore now assume the following.

Assumption A3. Assume $\mathrm{Cov}(\mathbf{X}) = \mathbf{I} = \mathrm{Cov}(\mathbf{S})$, and hence $\mathbf{A}$ to be orthonormal.

3.3. Uniqueness of the non-Gaussian decomposition

Theorem 3.2. *Assume A1–A3. Then there is a permutation* π *of* $\{1, \ldots, M\}$ *and some index* $1 \leq k < M$ *such that* $\mathbf{X}_1 \sim (\mathbf{S}_{\pi(1)}, \ldots, \mathbf{S}_{\pi(k)})$ *and* $\mathbf{X}_2 \sim (\mathbf{S}_{\pi(k+1)}, \ldots, \mathbf{S}_{\pi(M)})$.

Uniqueness of ISA can easily be established using this theorem, and most of the rest of this section will be devoted to its proof. Before proving it, we will show how it implies uniqueness of ISA.

Assumption A4. Assume $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_L)$ with irreducible, independent $\mathbf{X}_j$.

Again, similarly to above, assuming (A1) and (A4), decorrelation of all $\mathbf{X}_j$ and $\mathbf{S}_k$ may be performed without any loss of generality, so Assumption A3 does not conflict with these in any way, and given any combination of (A1), (A2) and (A4) we may always additionally assume (A3) (decorrelation of the independent components and orthonormality of the mixing matrix).

Theorem 3.3. *Assume A1–A4. Then* $L = M$ *and for every* $1 \leq k \leq M$ *there is some* $1 \leq j \leq L$ *such that* $\mathbf{S}_k \sim \mathbf{X}_j$.

Proof. We choose the component of minimal size among $\mathbf{S}_k$ and $\mathbf{X}_j$. Without loss of generality, we may assume this to be $\mathbf{X}_1$. Let us gather the other components of $\mathbf{X}$ to a single random vector:

$$\mathbf{X}' := (\mathbf{X}_2, \ldots, \mathbf{X}_L).$$

Then $(\mathbf{X}_1, \mathbf{X}')$ is an independent (but not irreducible if $L > 2$) decomposition and according to Theorem 3.2, $\mathbf{X}_1$ is equivalent to a combination of some $\mathbf{S}_k$. As $\mathbf{X}_1$ is the smallest component in the $\mathbf{S}_k$ and $\mathbf{X}_j$, it has to be equivalent to a single $\mathbf{S}_k$ for some k, and $\mathbf{X}'$ is equivalent to the concatenation of the other components of $\mathbf{S}$:

$$\mathbf{X}' \sim (\mathbf{S}_1, \ldots, \mathbf{S}_{k-1}, \mathbf{S}_{k+1}, \ldots, \mathbf{S}_M).$$

We remove $\mathbf{S}_k$ and $\mathbf{X}_1$ and proceed iteratively to get $M = L$, and for every k some j such that $\mathbf{S}_k \sim \mathbf{X}_j$. □

Let us now prove Theorem 3.2. We split up $\mathbf{A}$ into submatrices $\mathbf{A}_{jk}$ of size $\dim(\mathbf{X}_j) \times \dim(\mathbf{S}_k)$, so

$$\mathbf{X}_j = \sum_{k=1}^{M} \mathbf{A}_{jk}\mathbf{S}_k \tag{2}$$

and, equivalently,

$$\mathbf{S}_k = \sum_{j=1}^{2} \mathbf{A}_{jk}^\top \mathbf{X}_j \tag{3}$$

since $\mathbf{A}$ is orthonormal. The claim of Theorem 3.2 is equivalent to the claim that in every pair of matrices $\{\mathbf{A}_{1k}, \mathbf{A}_{2k}\}$ one of the two is zero. It suffices to show this for $k = 1$ as the proofs for the other cases are fully analogous. As $\mathbf{A}$ has full rank, $\operatorname{rank}(\mathbf{A}_{11}) + \operatorname{rank}(\mathbf{A}_{21}) \geq \dim(\mathbf{S}_1)$. If we have equality in this relation, we can easily show that both $\mathbf{A}_{11}$ and $\mathbf{A}_{21}$ being non-zero implies that $\mathbf{S}_1$ is reducible.

Lemma 3.3. *Assume a random vector $\mathbf{S}_1$ and two non-zero matrices $\mathbf{A}_1, \mathbf{A}_2$ such that $\operatorname{rank}(\mathbf{A}_1) + \operatorname{rank}(\mathbf{A}_2) = \dim(\mathbf{S}_1) = \operatorname{rank}(\mathbf{A}_1^\top \ \mathbf{A}_2^\top)$. If we can write*

$$\mathbf{S}_1 = (\mathbf{A}_1^\top \ \mathbf{A}_2^\top) \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

with independent random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, then $\mathbf{S}_1$ is reducible.

Proof. Let $D := \dim(\mathbf{S}_1)$ and $d := \dim\big(\ker(\mathbf{A}_1)\big)$. Using the rank-nullity theorem twice, we have

$$\begin{aligned} \dim\big(\ker(\mathbf{A}_2)\big) &= \dim(\mathbf{S}_1) - \operatorname{rank}(\mathbf{A}_2) = \operatorname{rank}(\mathbf{A}_1) \\ &= \dim(\mathbf{S}_1) - \dim(\ker(\mathbf{A}_1)) = D - d \end{aligned}$$

and we can then find a linearly independent set $\{\mathbf{v}_1, \ldots, \mathbf{v}_{D-d}\}$ such that $\mathbf{A}_2\mathbf{v}_j = 0$ for any $1 \leq j \leq D - d$. We also can find a linearly independent set $\{\mathbf{v}_{D-d+1}, \ldots, \mathbf{v}_D\}$ such that $\mathbf{A}_1\mathbf{v}_i = 0$ for any $D - d + 1 \leq i \leq D$. These two sets are guaranteed to be linearly independent, as $\operatorname{rank}(\mathbf{A}_1^\top \ \mathbf{A}_2^\top) = \dim(\mathbf{S}_1)$ and as $\mathbf{A}_1, \mathbf{A}_2$ were assumed to be non-zero, neither set is empty. Using these vectors, we define

$$\mathbf{B} := \begin{pmatrix} \mathbf{v}_1^\top \\ \vdots \\ \mathbf{v}_D^\top \end{pmatrix}.$$

Then

$$\begin{aligned} \mathbf{B}\mathbf{S}_1 &= (\mathbf{B}\mathbf{A}_1^\top \ \mathbf{B}\mathbf{A}_2^\top) \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{B}_1 & 0 \\ 0 & \mathbf{B}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{B}_1\mathbf{X}_1 \\ \mathbf{B}_2\mathbf{X}_2 \end{pmatrix} \end{aligned}$$

with some full rank matrices $\mathbf{B}_1$ and $\mathbf{B}_2$. It follows that $\mathbf{S}_1$ is reducible, as $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent and $\mathbf{B}$ is invertible. □

The other case contradicts the assumption that $\mathbf{S}$ contains no independent, normally distributed component.

Lemma 3.4. *Assume A1–A3. If $\operatorname{rank}(\mathbf{A}_{11}) + \operatorname{rank}(\mathbf{A}_{21}) > \dim(\mathbf{S}_1)$, then $\mathbf{S}_1$ contains an independent Gaussian.*

Note that this does not necessarily imply reducibility of $\mathbf{S}_1$, as it might simply be just a one-dimensional Gaussian. In order to prove this claim, we need the following technical lemma, the proof of which we have moved to the Appendix.

Lemma 3.5. *Assume $L, M \in \mathbb{N}$, functions*

$$f_k : \mathbb{R}^{m_k} \to \mathbb{C} \quad (k = 1, \ldots, L)$$

and

$$g_j : \mathbb{R}^{n_j} \to \mathbb{C} \quad (j = 1, \ldots, M)$$

and matrices $\mathbf{A}_{jk} : \mathbb{R}^{m_k} \to \mathbb{R}^{n_j}$ where $\sum_{k=1}^{L} m_k = \sum_{j=1}^{M} n_j$ such that

(i) *the functions f_k, g_j are twice continuously differentiable*
(ii) *the matrix $\mathbf{A} := (\mathbf{A}_{ij})_{i,j}$ is invertible*

(iii) *for any* $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_L) \in (\mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_L})$

$$\prod_{k=1}^{L} f_k(\mathbf{x}_k) = \prod_{j=1}^{M} g_j(\mathbf{A}_j \mathbf{x}) \tag{4}$$

where $\mathbf{A}_j = (\mathbf{A}_{j1} \dots \mathbf{A}_{jL})$.

Then, for any two indices $1 \leq i < j \leq L$ *and for any* M*-tuple of points* $(\mathbf{y}_1, \dots, \mathbf{y}_M) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_M}$ *where* $g_k(\mathbf{y}_k) \neq 0$ $(k = 1, \dots, M)$*, locally*

$$\sum_{k=1}^{M} \mathbf{A}_{ki}^{\top} (\mathbf{H}_{\log(g_k)}|_{\mathbf{y}_k}) \mathbf{A}_{kj} = 0$$

where log *is a (local) complex logarithm and* $\mathbf{H}_f$ *denotes the Hessian of* f.

Using this, we can now prove Lemma 3.4.

Proof. Let us define $D := \dim(\mathbf{S}_1)$, and $d_i := \dim(\ker(\mathbf{A}_{i1}))$ $(i = 1, 2)$. Then $d_1 + d_2 = (D - \operatorname{rank}(\mathbf{A}_{11})) + (D - \operatorname{rank}(\mathbf{A}_{21})) < D$. We then choose vectors $\mathbf{v}_2, \dots, \mathbf{v}_{d_1+1}$ that form a basis of $\ker(\mathbf{A}_{11})$ and similarly $\mathbf{v}_{D-d_2+1}, \dots, \mathbf{v}_D$ (note here that $D - d_2 + 1 > d_1 + 1$) that form a basis of $\ker(\mathbf{A}_{21})$. As the matrix $\mathbf{A}$ is invertible, $\ker(\mathbf{A}_{11})$ and $\ker(\mathbf{A}_{21})$ are disjoint, so the set of vectors $\{\mathbf{v}_2, \dots, \mathbf{v}_{d_1+1}, \mathbf{v}_{D-d_2+1}, \dots, \mathbf{v}_D\}$ is linearly independent. Now choose vectors $\mathbf{v}_k$ $(k = d_1 + 2, \dots, D - d_2)$ – this set might be empty – such that the vectors $\mathbf{v}_2, \dots, \mathbf{v}_D$ are linearly independent, and finally choose a $\mathbf{v}_1$ orthogonal to $\operatorname{span}(\{\mathbf{v}_2, \dots, \mathbf{v}_D\})$. We define $\mathbf{T}_0 := (\mathbf{v}_1, \dots, \mathbf{v}_D)$, and then

$$\mathbf{T} := \begin{pmatrix} \mathbf{T}_0 & 0 \\ 0 & \mathbf{I} \end{pmatrix},$$

where $\mathbf{I}$ is the $(\dim(\mathbf{S}) - D)$-dimensional identity. The first column of $\mathbf{T}$ is orthogonal to the other columns, so the first row of $\mathbf{T}^{-1}$ is orthogonal to the other rows of $\mathbf{T}^{-1}$. Now

$$\mathbf{X} = \mathbf{AS} = \mathbf{AT}(\mathbf{T}^{-1}\mathbf{S})$$

where $\mathbf{T}^{-1}$ is an operation purely within $\mathbf{S}_1$. Therefore, we may replace $\mathbf{A}$ with $\mathbf{AT}$ and $\mathbf{S}$ with $\mathbf{T}^{-1}\mathbf{S}$. Note that due to our choice of $\mathbf{T}$ we do not have full decorrelation of $\mathbf{S}_1$ anymore; only $(\mathbf{S}_1)_1$ is decorrelated from the other components of $\mathbf{S}_1$ due to the first row of the transformation $\mathbf{T}^{-1}$ being orthogonal to its other rows. Now, the columns of $\mathbf{A}_{11}$ with indices $2, \dots, d_1 + 1$ contain only zeros, and so do the columns of $\mathbf{A}_{21}$ with indices $D - d_2 + 1, \dots, D$. The other columns of $\mathbf{A}_{11}$ have full rank, and so do the other columns of $\mathbf{A}_{21}$.

Let us now turn to the characteristic functions of $\mathbf{X}$ and $\mathbf{S}$. Due to their independent decompositions, we have

$$\prod_{k=1}^{2} \widehat{\mathbf{X}}_k(\mathbf{x}_k) = \widehat{\mathbf{X}}(\mathbf{x}) = \widehat{\mathbf{AS}}(\mathbf{x}) = \widehat{\mathbf{S}}(\mathbf{A}^{\top}\mathbf{x}) = \prod_{j=1}^{M} \widehat{\mathbf{S}}_j(\mathbf{A}_j^{\top}\mathbf{x}).$$

For now we fix an $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_M)$ where $\widehat{\mathbf{S}}_k(\mathbf{A}_k^{\top}\mathbf{x}_k) \neq 0$ (this exists as $\widehat{\mathbf{S}}(0) = 1$). As $\mathbf{S}$ and $\mathbf{X}$ have existing covariance, their characteristic functions are twice continuously differentiable, so all assumptions of Lemma 3.5 are fulfilled; therefore, locally

$$0 = \sum_{k=1}^{M} \mathbf{A}_{1k} \mathbf{H}_{\mathbf{h}_k}|_{\mathbf{s}_k} \mathbf{A}_{2k}^{\top}$$

where $\mathbf{s}_k := \mathbf{A}_k\mathbf{x}$ and $\mathbf{h}_k := \log(\widehat{\mathbf{S}}_k)$ is a local logarithm. In this sum, every summand depends on a different variable $\mathbf{s}_k$, therefore every summand is constant: for every k

$$\mathbf{A}_{1k} \mathbf{H}_{\mathbf{h}_k}|_{\mathbf{s}_k} \mathbf{A}_{2k}^{\top} = \mathbf{C}_k \tag{5}$$

with some constant matrices $\mathbf{C}_k$, and in particular the first summand is constant:

$$\mathbf{A}_{11} \mathbf{H}_{\mathbf{h}_1}|_{\mathbf{s}_1} \mathbf{A}_{21}^{\top} = \mathbf{C}. \tag{6}$$

Now, there are invertible matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ such that the first $D - d_2$ rows of $\mathbf{B}_2\mathbf{A}_{21}$ consist of the first $D - d_2$ unit vectors, and such that the k-th row of $\mathbf{B}_1\mathbf{A}_{11}$ consists of the k-th unit vector, where $k = 1, d_1 + 2, \dots, D$:

$$\mathbf{e}_j^{\top} \mathbf{B}_2 \mathbf{A}_{21} = \mathbf{e}_j^{\top} \quad (j = 1, \dots, D - d_2);$$
$$\mathbf{e}_j^{\top} \mathbf{B}_2 \mathbf{A}_{21} = 0 \quad (j = D - d_2 + 1, \dots, D)$$

and

$$\mathbf{e}_k^\top \mathbf{B}_1\mathbf{A}_{11} = \mathbf{e}_k^\top \quad (k = 1, d_1+2, \ldots, D);$$
$$\mathbf{e}_k^\top \mathbf{B}_1\mathbf{A}_{11} = 0 \quad (k = 2, \ldots, d_1+1).$$

Multiplying Eq. (6) with $\mathbf{B}_1$ from the left and $\mathbf{B}_2^\top$ from the right gives us

$$\mathbf{B}_1\mathbf{A}_{11}\mathbf{H}_{h_1}|_{\mathbf{s}_1}(\mathbf{B}_2\mathbf{A}_{21})^\top = \mathbf{C} \tag{7}$$

for some matrix $\mathbf{C}$. Multiplication of Eq. (7) first with $\mathbf{e}_1^\top$ from the left and $\mathbf{e}_j$ ($j = 1, \ldots, D-d_2$) from the right gives us

$$\partial_1\partial_j h_1(\mathbf{s}_1) = \mathbf{e}_1^\top \mathbf{H}_{h_1}|_{\mathbf{s}_1}\mathbf{e}_j = c_j$$

with some constants c_j and $j \in \{1, \ldots, D-d_2\}$. Similarly, multiplication of Eq. (7) with $\mathbf{e}_1$ from the right and $\mathbf{e}_k^\top$ ($k = d_1+2, \ldots, D$) from the left gives us

$$\partial_1\partial_k h_1(\mathbf{s}_1) = \partial_k\partial_1 h_1(\mathbf{s}_1) = \mathbf{e}_k^\top \mathbf{H}_{h_1}|_{\mathbf{s}_1}\mathbf{e}_1 = c_k$$

with some constants c_k and $k \in \{d_1+2, \ldots, D\}$. We have assumed $d_1 < D-d_2$, so also $d_1+2 \leq D-d_2+1$, so all in all

$$\partial_1\partial_k h_1(\mathbf{s}_1) = c_k$$

for all $1 \leq k \leq D$. Integrating the k-th such equation by s_k tells us that

$$\partial_1 h_1(\mathbf{s}_1) = c_k s_k + g(s_1, \ldots, s_{k-1}, s_{k+1}, \ldots, s_D)$$

for some g independent of s_k, and so together $\partial_1 h_1(\mathbf{s}_1) = \sum_{k=1}^D c_k s_k + C$ with some constant C. Integration then shows

$$h_1(\mathbf{s}_1) = \sum_{k=1}^D c_k s_k s_1 + c_0 s_1 + g(s_2, \ldots, s_D)$$

(note that we have redefined c_1 here) with some continuous function g that does not depend on s_1, and we then get

$$\widehat{\mathbf{S}}_1(\mathbf{s}_1) = \exp\left(\sum_{k=1}^D c_k s_1 s_k + c_0 s_1 + g(s_2, \ldots, s_D)\right).$$

Such a representation exists in a neighborhood of any $\mathbf{s}$ where $\widehat{\mathbf{S}}(\mathbf{s}) \neq 0$. This is an open condition, so the set of all such $\mathbf{s}$ is open again. But due to continuity the representation above also holds in the closure of this set, hence in a clopen set. As $\widehat{\mathbf{S}}(0) = 1 \neq 0$, this set cannot be empty, therefore this has to hold everywhere.

Let us now use what we know on the first two moments (expectation and covariance) to infer some information on the constants c_k. As $\mathbf{S}_1$ is centered and we know the first row (and column) of its covariance matrix, let us calculate the according expressions in terms of $\widehat{\mathbf{S}}_1$ using Eq. (1) from Section 2.3:

$$\frac{\partial}{\partial s_1}\widehat{\mathbf{S}}_1 = \widehat{\mathbf{S}}_1(\mathbf{s}_1)\left(2c_1 s_1 + \sum_{k=2}^D c_k s_k + c_0\right)$$

so plugging in $\mathbf{s} = 0$, we see $c_0 = 0$. Next, for $j \neq 1$, we have

$$\frac{\partial}{\partial s_j}\widehat{\mathbf{S}}_1 = \widehat{\mathbf{S}}_1(\mathbf{s}_1)\left(c_j s_1 + \frac{\partial}{\partial s_j} g(s_2, \ldots, s_D)\right)$$

so, again plugging in $\mathbf{s} = 0$, we get $\frac{\partial}{\partial s_j} g(s_2, \ldots, s_D)|_{\mathbf{s}=0} = 0$ for all $j \neq 1$. Now we use the fact that $E\{\mathbf{S}_{11}\mathbf{S}_{1j}\} = 0$ for all $j \neq 1$

$$\frac{\partial^2}{\partial s_1 \partial s_j}\widehat{\mathbf{S}}_1 = \widehat{\mathbf{S}}_1(\mathbf{s}_1)\left(c_j s_1 + \frac{\partial}{\partial s_j} g(s_2, \ldots, s_D)\right)\left(2c_1 s_1 + \sum_{k=2}^D c_k s_k\right) + c_j \widehat{\mathbf{S}}_1(\mathbf{s}_1)$$

so, plugging in $\mathbf{s} = 0$, we get

$$0 = c_j$$

so $c_j = 0$ for all $j \neq 1$. Altogether, we have

$$\widehat{\mathbf{S}}_1(\mathbf{s}_1) = \exp(c_1 s_1^2 + g(s_2, \ldots, s_D)) = \exp(c_1 s_1^2)\exp(g(s_2, \ldots, s_D)).$$

As $\widehat{\mathbf{S}}_1$ factorizes this way, the first component of $\mathbf{S}_1$ is a Gaussian independent of the rest. For the sake of completeness, similarly to above, $\mathbf{S}_1$ being white implies c_2 to be $-1/2$, but for our proof it suffices to notice that the first component of $\mathbf{S}_1$ is a Gaussian independent of the rest of $\mathbf{S}_1$, which is what we wanted to show. □

Let us summarize the proof of Theorem 3.2.

Proof. The claim is equivalent to the fact that in Eq. (2), for every k, one of the two $\mathbf{A}_{k1}$ and $\mathbf{A}_{k2}$ is zero, or, equivalently, $\operatorname{rank}(\mathbf{A}_{k1}) = 0$ or $\operatorname{rank}(\mathbf{A}_{k2}) = 0$. We have $\operatorname{rank}\mathbf{A}_{k1}, \operatorname{rank}\mathbf{A}_{k2} \leq \dim(\mathbf{S}_k)$, and as $\mathbf{A}$ is invertible, $\operatorname{rank}(\mathbf{A}_{k1}) + \operatorname{rank}(\mathbf{A}_{k2}) \geq \dim(\mathbf{S}_k)$.

Assume first $\operatorname{rank}(\mathbf{A}_{k1}) + \operatorname{rank}(\mathbf{A}_{k2}) = \dim(\mathbf{S}_k)$. Eq. (3) tells us that $\mathbf{S}_k = \mathbf{A}_{k1}^\top \mathbf{X}_1 + \mathbf{A}_{k1}^\top \mathbf{X}_2$ and now we are facing exactly the assumptions of Lemma 3.3 so $\mathbf{S}_k$ is reducible, contradicting the assumption of irreducibility of all $\mathbf{S}_k$.

If $\operatorname{rank}(\mathbf{A}_{k1}) + \operatorname{rank}(\mathbf{A}_{k2}) > \dim(\mathbf{S}_k)$, all assumptions of Lemma 3.4 are fulfilled, so $\mathbf{S}_k$ contains (or is) an independent Gaussian, contradicting the assumption of non-Gaussianity of $\mathbf{X}$. □

Finally, collecting Theorems 3.3 and 3.1, we conclude with the uniqueness of ISA in general.

Theorem 3.4. *The linear decomposition of a random vector* $\mathbf{X}$ *with existing covariance into independent, irreducible subspaces is unique up to the order of the components, invertible transformations purely within the total Gaussian subspace and invertible transformations within the non-Gaussian subspaces.*

In other words, this means that the set of vector subspaces $\mathbf{X}$ is projected to is unique, but for any vector subspace, we may freely choose a basis.

4. An ISA algorithm via joint block diagonalization of the Hessian

The key idea in the proof of Theorem 3.4 was block diagonality of the Hessian of the second characteristic function of $\mathbf{S}$. This naturally motivates an algorithm that performs ISA by recovering this block diagonality structure. In the setting of ICA (in which case the Hessian even is diagonal everywhere) a similar, simpler approach has been proposed in [39]. A high level description of the algorithm is given in Algorithm 1, with the remainder of this section devoted to a more detailed explanation of the parts of the algorithm.

Input: d dimensional real random vector $\mathbf{X}$ with $\operatorname{Cov}(\mathbf{X}) = \mathbf{I}_d$
Output: $\mathbf{W} \in O(d)$ and tuple of integers $(d_1, \ldots, d_N)$ such that $(\mathbf{S}_1^\top, \ldots, \mathbf{S}_N^\top)^\top := \mathbf{S} := \mathbf{WX}$ with irreducible, mutually independent $\mathbf{S}_k$, where $\dim(\mathbf{S}_k) := d_k$.

$\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \leftarrow k$ randomly chosen points in $\mathbb{R}^d$;
$\mathbf{M}_k \leftarrow \mathbf{H}_{\log \mathbf{X}}|_{\mathbf{x}_k}$; `// Subsection 4.4`
$\mathbf{W} \leftarrow \mathtt{rJBD}(\{\mathbf{M}_1, \ldots, \mathbf{M}_k\})$; `// ℝ-JBD, Subsection 4.3`
for *i=1* **to** k **do** $\mathbf{M}_i \leftarrow \mathbf{W}^\top \mathbf{M}_i \mathbf{W}$; `// now Mᵢ are block-diagonal`
$\mathcal{M} \leftarrow \{\mathbf{M}_1, \ldots, \mathbf{M}_k\}$;
$(d_1, \ldots, d_N) \leftarrow \mathtt{comps}(\mathcal{M})$; `// recover components, Subsection 4.5`

Algorithm 1: `HessianISA` (high level description)

4.1. ISA via joint block diagonalization

The proof of Theorem 3.4 is based on the one-to-one correspondence between the (mutually) independent subspaces of a random vector $\mathbf{S}$ and the block diagonality structure of the Hessian of its second characteristic function: if $\mathbf{S} = (\mathbf{S}_1^\top, \ldots, \mathbf{S}_N^\top)^\top$ where the subspaces $\mathbf{S}_j$ are mutually independent, then $\mathbf{H}_{\widehat{\mathbf{S}}}$ is block diagonal everywhere with blocks of size $\dim(\mathbf{S}_1), \ldots, \dim(\mathbf{S}_N)$. The Hessian of the second characteristic function of a d-dimensional random vector $\mathbf{S}$ transforms as follows under linear transformations $\mathbf{A} \in \mathrm{Gl}(d, \mathbb{R})$

$$\mathbf{H}_{\log \widehat{\mathbf{AS}}}|_{\mathbf{x}} = \mathbf{A}\mathbf{H}_{\log \widehat{\mathbf{S}}}|_{\mathbf{A}^\top \mathbf{x}} \mathbf{A}^\top \tag{8}$$

for any $\mathbf{x} \in \mathbb{R}^d$ where it exists, that is, where $\widehat{\mathbf{AS}}(\mathbf{x}) \neq 0$. This follows directly by evaluating part (ii) of Lemma 2.2 for every $1 \leq i, j \leq d$. These facts can be used to recover the irreducible, independent subspaces in a random vector that is only observed after being mixed in an invertible, linear way – or, equivalently, to decompose an arbitrary random vector $\mathbf{X}$ into irreducible, independent subspaces.

Eq. (8) allows us to reformulate this as a *Joint Block Diagonalization* (JBD) task. For this let us assume a partition $\mathbf{d} = (d_1, \ldots, d_N)$ of the integer d, i.e. we write d as the sum of positive integers, $d = d_1 + \cdots + d_N$, where the order of the summands does not matter and without loss of generality we may assume them to be ordered in descending order, $d_1 \geq d_2 \geq \cdots \geq d_N$. Then a $d \times d$ matrix $\mathbf{M}$ is said to be $\mathbf{d}$-blockdiagonal if it is of the form

$$\begin{pmatrix} \mathbf{M}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{M}_N \end{pmatrix} \tag{9}$$

where $\mathbf{M}_j$ is some square $d_j \times d_j$ matrix, and a set $\mathcal{M}$ of $d \times d$ matrices is said to be jointly $\mathbf{d}$-blockdiagonal (or $\mathbf{d}$-JBD) if every $\mathbf{M} \in \mathcal{M}$ is. If we do not care for the partition, we simply say that a matrix $\mathbf{M}$ is block-diagonal, resp. that the set of matrices $\mathcal{M}$ has JBD form. Assume for the moment a fixed partition $\mathbf{d}$ of d. Given now an arbitrary set $\mathcal{M}$ of matrices of size $d \times d$, the task of JBD is to find a transformation $\mathbf{W}$ of size $d \times d$ such that $\mathbf{W}^{-1}\mathcal{M}\mathbf{W} = \{\mathbf{W}^{-1}\mathbf{M}\mathbf{W} : \mathbf{M} \in \mathcal{M}\}$ is jointly $\mathbf{d}$-blockdiagonal. If the entries of the matrices in $\mathcal{M}$ are all real-valued, the choice of $\mathbf{W}$ is restricted to $O(d)$, the group of orthogonal matrices. If the entries of the matrices in $\mathcal{M}$ are complex-valued, $\mathbf{W}$ is restricted to be unitary, i.e. $\mathbf{W}$ has to fulfill $\mathbf{W}\mathbf{W}^* = \mathbf{I}_d$, where $\mathbf{W}^*$ denotes the Hermitian adjoint (or complex transpose) of $\mathbf{W}$. Obviously, if $\mathbf{W}^{-1}\mathcal{M}\mathbf{W}$ is jointly $\mathbf{d}$-blockdiagonal, then so is $(\mathbf{W}\mathbf{T})^{-1}\mathcal{M}(\mathbf{W}\mathbf{T})$ whenever $\mathbf{T}$ is a $\mathbf{d}$-blockdiagonal orthogonal (resp. unitary) $d \times d$ matrix. Dropping the order of the partition $\mathbf{d}$, a second indeterminacy is given by permutation of whole blocks, corresponding to a multiplication from the right with a block-permutation matrix. Just as with ISA, these two indeterminacies exist for *any* solution to the JBD task, and can therefore be seen as *global* indeterminacies. An evident additional indeterminacy lies in the partition $\mathbf{d}$: in ISA we were not only interested in a decomposition being independent, but we also wanted it to be the finest such one, namely one where the single subspaces were irreducible. With JBD, we have a similar setting. If $\mathbf{d}'$ is another partition of d that is coarser than $\mathbf{d}$ (i.e. if it can be expressed merely by grouping together some of the components of $\mathbf{d}$, e.g. if $\mathbf{d} = (4, 3, 2)$ and $\mathbf{d}' = (7, 2)$ or $\mathbf{d}' = (6, 3)$ or even $\mathbf{d}' = (9)$), then any $\mathbf{d}$-JBD set of matrices is also $\mathbf{d}'$-JBD. Even more so, in general any set of $d \times d$ matrices will always be jointly (d)-JBD, i.e. consist of one large $d \times d$ block. Clearly, a finer decomposition is always preferred over a coarser one, so we demand that $\mathbf{d}$ have maximal possible length, corresponding to a decomposition with the highest possible number of blocks (in which case the blocks cannot be decomposed further with linear means and then are called *irreducible*). While this is not obvious, one can show that then the sizes of the blocks are unique. Then a matrix $\mathbf{W}$ (orthogonal or unitary, depending on the setting) is said to solve the JBD-task of $\mathcal{M}$ if $\mathbf{W}^{-1}\mathcal{M}\mathbf{W}$ is jointly $\mathbf{d}$-blockdiagonal where $\mathbf{d}$ has maximal possible length.

In general, evaluation of the left hand side of Eq. (8) results in matrices with complex entries, giving rise to complex JBD. However our mixing matrix $\mathbf{A}$ was assumed to have only real entries, and by whitening of the sources to be orthogonal; hence the search space for $\mathbf{W}$ should also be restricted to only real (and thus orthogonal) matrices. We therefore transform our setting into the task to JBD of real matrices by not performing JBD on $\mathcal{M}$ itself, but by performing JBD on the set $\mathcal{M}' := \mathrm{Re}(\mathcal{M}) \cup \mathrm{Im}(\mathcal{M})$, where Re(.) (resp. Im(.)) is the operation taking the real (resp. the imaginary) part of its argument.

4.2. A local indeterminacy of JBD

It is interesting to note that JBD has one further indeterminacy, see e.g. [22] but this can only happen if the matrices in $\mathcal{M}$ have additional structure, namely if two (or more) of the irreducible blocks form a so-called *simple* block. Assuming a (minimal) $\mathbf{d}$-blockdiagonal representation of $\mathcal{M}$ (i.e. any $\mathbf{M} \in \mathcal{M}$ has the form (9)), the i-th and the j-th block of $\mathcal{M}$ lie in a single simple block if and only if there is a single orthogonal (resp. unitary) $\mathbf{W}_{ij}$ such that for every $\mathbf{M} \in \mathcal{M}$ the i-th and the j-th block are connected via the relation $\mathbf{M}_j = \mathbf{W}_{ij}^{-1}\mathbf{M}_i\mathbf{W}_{ij}$. A simple block then consists of all irreducible blocks connected this way. Non-trivial simple blocks (i.e. simple blocks consisting of two or more irreducible blocks) can occur only for *some* (sets of) matrices, thus we see it as only a *local* indeterminacy. As we already know that ISA has only two global indeterminacies, this will never occur when performing JBD on Hessians of the second characteristic function of a random vector, so the only indeterminacies of the JBD-problem in our setting are permutation of blocks and invertible transformations within blocks—exactly the same as the indeterminacies of ISA. Therefore, recovery of $\mathbf{S}$ given only $\mathbf{X} = \mathbf{A}\mathbf{S}$ is possible via JBD as follows: let $\mathcal{M}$ be the left hand side of Eq. (8), evaluated for a number of evaluation points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. If $\mathbf{W}^{-1}\mathcal{M}\mathbf{W}$ has JBD form for some orthogonal $d \times d$ matrix $\mathbf{W}$ with real entries, then $\mathbf{A} = \mathbf{W}$ (up to the global indeterminacies). Note that while it is possible that a particularly bad choice of evaluation points may give us a JBD form that is finer than the decomposition of $\mathbf{X}$ itself (for example, at $\mathbf{x} = 0$, the left hand side of Eq. (8) simply is $-\mathrm{Cov}(\mathbf{X})$), every dependency between some components of $\mathbf{S}$ always is reflected by non-zero entries in the Hessian for *some* evaluation points.

4.3. JBD via joint diagonalization

In general, the exact JBD criterion of the entries outside of the block-diagonal being exactly 0 usually is not fulfilled in real-world situations. In our case we have matrices in theoretically exact JBD form on the right hand side of Eq. (8). In practice, however often only a finite number of samples of $\mathbf{S}$ is given, resulting at least in a finite-sample estimation error of $\mathbf{S}$ itself (and then $\widehat{\mathbf{S}}$), and then also in an error in the estimation of the Hessian of $\log \widehat{\mathbf{S}}$ (which will be discussed below). Further estimation errors may be given if $\mathbf{X}$ contains additional noise, i.e. instead of our exact setting only $\mathbf{X} = \mathbf{A}\mathbf{S} + \varepsilon$ holds where ε is some kind of noise model. It is therefore convenient not to seek an exact JBD solution but rather allow off-diagonal entries, as long as their absolute value is below some threshold τ. The choice of such a threshold τ is non-trivial; therefore we instead propose simply applying a Joint Diagonalization (JD) algorithm to the set $\mathcal{M}'$.

Given a set of square matrices $\mathcal{B}$ with real entries, a JD algorithm searches for an orthogonal matrix $\mathbf{W}$ such that the set $\mathbf{W}^\top\mathcal{B}\mathbf{W}$ is "as diagonal as possible". Formally, if $\mathcal{B} = \{\mathbf{M}_1, \ldots, \mathbf{M}_N\}$, the JD algorithm minimizes the cost function

$$\sum_{k=1}^{N} \mathrm{off}(\mathbf{W}^\top\mathbf{M}_k\mathbf{W})$$

among all $\mathbf{W} \in O(d)$, where off(.) calculates the Frobenius norm (i.e. the sum of the squares) of the off-diagonal entries of its argument. Instead of minimizing the Frobenius norm of the off-diagonal entries, one can equivalently maximize the Frobenius norm of the diagonal entries in the sum. In many signal processing situations it is hypothesized that any JD solution already is equivalent to a JBD solution apart from a final permutation step; see for example [1,36]. In general this however is not true, as we can see in the following simple setting, where $\mathcal{S}$ consists of only a single matrix.

Example 4.1. Let

$$\mathbf{M} := \begin{pmatrix} 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & 1 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 \end{pmatrix}.$$

Then $\operatorname{off}(\mathbf{W}^\top \mathbf{M}\mathbf{W}) = 0$ for any $\mathbf{W} \in O(4)$, so the JD cost function is constant under all $\mathbf{W} \in O(4)$. But if

$$\mathbf{W} = \frac{1}{2}\begin{pmatrix} 1 & -1 & 1 & 1 \\ 0 & \sqrt{2} & 0 & \sqrt{2} \\ -1 & -1 & -1 & 1 \\ \sqrt{2} & 0 & -\sqrt{2} & 0 \end{pmatrix} \in O(4)$$

then $\mathbf{W}^\top \mathbf{M}\mathbf{W}$ is $(2, 2)$-blockdiagonal, showing that there indeed is a better blockdiagonal decomposition than one suggested by JD.

The reason why here JD does not provide a JBD solution lies in the fact that $\mathbf{M} = -\mathbf{M}^\top$, i.e. $\mathbf{M}$ is anti-symmetric, and then so is $\mathbf{W}^\top \mathbf{M}\mathbf{W}$ for any $\mathbf{W} \in O(d)$. Therefore, the JD cost function is constant under arbitrary additions of anti-symmetric matrices. Intuitively, it only "sees" the symmetric part of its arguments. In our case this is sufficient, as every $\mathbf{M}$ in $\mathcal{M}$ is symmetric (as it is a Hessian matrix, which always is symmetric) and so then are the matrices in $\mathcal{M}'$. The advantage over a full JBD algorithm is the lack of the threshold parameter τ, and we argue that the conjecture of performing JBD via JD and then seeking a final recovery permutation is valid in our setting.

4.4. Estimating the Hessian

We still have to calculate the left hand side of Eq. (8) for our evaluation points $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. Simple rewriting shows that whenever $\widehat{\mathbf{X}}(\mathbf{x}) \neq 0$, we can rewrite the Hessian of the second characteristic function of a random vector $\mathbf{X}$ as

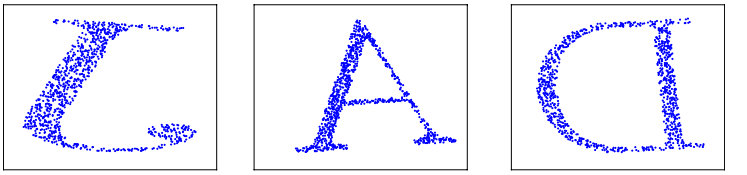
$$\begin{aligned} \mathbf{H}_{\log \widehat{\mathbf{X}}}|_{\mathbf{x}} &= \frac{\mathbf{H}_{\widehat{\mathbf{X}}}|_{\mathbf{x}}}{\widehat{\mathbf{X}}(\mathbf{x})} - \frac{\left(\mathbf{d}\widehat{\mathbf{X}}|_{\mathbf{x}}\right)^\top \mathbf{d}\widehat{\mathbf{X}}|_{\mathbf{x}}}{\left(\widehat{\mathbf{X}}(\mathbf{x})\right)^2} \\ &= \frac{E\{\exp(i\mathbf{X}^\top\mathbf{x})\mathbf{X}\}E\{\exp(i\mathbf{X}^\top\mathbf{x})\mathbf{X}^\top\}}{\left(E\{\exp(i\mathbf{X}^\top\mathbf{x})\}\right)^2} - \frac{E\{\exp(i\mathbf{X}^\top\mathbf{x})\mathbf{X}\mathbf{X}^\top\}}{E\{\exp(i\mathbf{X}^\top\mathbf{x})\}}. \end{aligned}$$

Given i.i.d. samples $\mathbf{X}_1, \ldots, \mathbf{X}_k$ of $\mathbf{X}$ a straight-forward estimator (not necessarily the best one) of $\mathbf{H}_{\log \widehat{\mathbf{X}}}|_{\mathbf{x}}$ therefore is given by

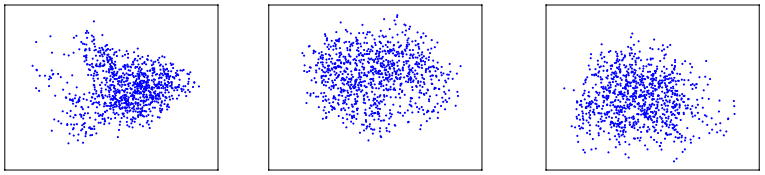
$$\frac{\sum_{j=1}^k \left(\exp(i\mathbf{X}_j^\top\mathbf{x})\mathbf{X}_j\right) \sum_{j=1}^k \left(\exp(i\mathbf{X}_j^\top\mathbf{x})\mathbf{X}_j^\top\right)}{\left(\sum_{j=1}^k \exp(i\mathbf{X}_j^\top\mathbf{x})\right)^2} - \frac{\sum_{j=1}^k \left(\exp(i\mathbf{X}_j^\top\mathbf{x})\mathbf{X}_j\mathbf{X}_j^\top\right)}{\sum_{j=1}^k \exp(i\mathbf{X}_j^\top\mathbf{x})}.$$

4.5. Recovering the permutation

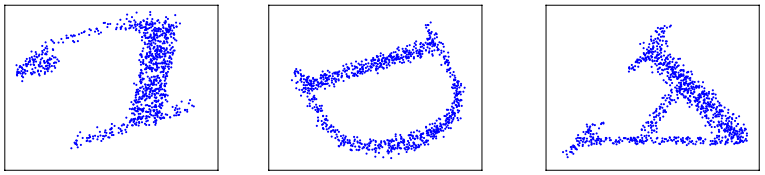
The JD algorithm returns an orthogonal $\mathbf{W}$ that makes the set of real and imaginary parts of the (numerical approximation of the) Hessian of $\mathbf{X}$ at the chosen evaluation points as diagonal as possible via an orthogonal transformation (rotation). This $\mathbf{W}$ then corresponds to $\mathbf{A}$ apart from a final permutation. In other words, $\mathbf{W}^\top\mathbf{X} = \mathbf{P}\mathbf{S}$ and all that is left to do now is finding $\mathbf{P}$. In order to simplify notation, we replace every $\mathbf{M} \in \mathcal{M}'$ with $\mathbf{W}^\top\mathbf{M}\mathbf{W}$, i.e. the matrices in $\mathcal{M}'$ already represent the demixed (although not yet grouped) sources. Let us illustrate how to recover the permutation assuming that $\mathcal{M}'$ perfectly fits the model, i.e. the off-blockdiagonal entries are exactly 0. In this case we only need to differ between entries being non-zero (indicating two components being connected) and 0 (which does not necessarily indicate two non-connected components, as this might simply be an artifact of a bad choice of a evaluation point, as explained above), and we therefore w.l.o.g. can assume every $\mathbf{M} \in \mathcal{M}'$ to have entries only in $\{0, 1\}$. We can then interpret the problem as the task of finding connected components on a graph consisting of d nodes [32], where there is an edge connecting nodes i and j (i.e. they are directly connected) if and only if $M_{ij} = 1$ for some $\mathbf{M} \in \mathcal{M}'$. Equivalently, we can calculate $\sum_{\mathbf{M}\in\mathcal{M}'} \mathbf{M}$, replace all non-zero entries with the value 1 and ask for the connected components of the graph corresponding to this adjacency matrix.



(a) Sources.



(b) Mixtures.



(c) Recoveries given by a typical run.

Fig. 1. Illustration of the data: (a) unmixed sources $\mathbf{S}$, (b) mixed observations $\mathbf{X} = \mathbf{AS}$, and (c) recovered subspaces $\mathbf{WX}$. For each of the three 6-dimensional signals, the plots depict the projections onto the subspace spanned by the first two (left), the third and fourth (middle) and the fifth and sixth (right) axes.

In the numerical (non-perfect) case, we may still perform a similar approach, just that we do not differ between zero and non-zero entries of the matrices $\mathbf{M}$, but rather see all entries whose absolute value lies below a threshold τ (the choice of which is non-trivial) as equivalent to 0, in other words, the i-th and the j-th node of our graph are directly connected if there is at least one $\mathbf{M}$ such that $|M_{ij}| > \tau$. By introducing a threshold parameter only at this stage and not already earlier, as it would have been the case with a direct JBD algorithm, we have lessened its potential negative impact on the outcome. In the worst case, we still get the demixed (1d-)components, but with no indication which of them represent a subspace and which are (mutually) independent.

4.6. Simulations

We illustrate the feasibility of the approach in a simulated setting. We generated $N = \{0.1, 0.5, 1, 2, \ldots, 10\} \times 10^3$ samples from three 2-dimensional sources, corresponding to letters of the alphabet (i.e. we uniformly sample pixels in the letters), which have been centered and normalized to have unit covariance; see Fig. 1(a). The $d = 6$-dimensional source $\mathbf{S}$ is then mixed according to $\mathbf{X} := \mathbf{AS}$ where $\mathbf{A}$ is a random (uniformly sampled according to the Haar measure) orthogonal matrix; see Fig. 1(b) for an illustration of the mixtures. The evaluation points are chosen such that every one contains exactly two non-zero entries, which are set to 1. We now calculate the empirical Hessians of $\log \widehat{\mathbf{X}}$ at $\mathbf{x}$ and perform real joint diagonalization of the set of real and imaginary parts of them. The demixed sources then are clustered into subspaces as described above, where the threshold parameter was set to $\tau = 0.11$, which was chosen by manual inspection. The reconstruction quality was quantified by calculating the Frobenius norm of the entries on the (2-2-2) blocks of $\mathbf{W}^\top\mathbf{A}$ where $\mathbf{W}$ denotes the recovery matrix (possibly after some blockwise permutation) and dividing this by $d = 6$ for normalization. We have perfect recovery if $\mathbf{W}^\top\mathbf{A}$, the product of recovery and mixing matrix is block-diagonal, in which case the reconstruction quality is 1. Empirically, for a randomly (uniformly, according to the Haar measure on $O(6)$) chosen recovery matrix, the reconstruction quality is about 0.48. Even in the theoretically worst case, at least for some blockwise permutations the blockwise Frobenius norm is positive; therefore even here we achieve positive recovery quality.

Example 4.2. In the case of $(2 + 2 + 2)$-dimensional sources, we see worst recovery if $\mathbf{W}^\top\mathbf{A}$ takes the form

$$\mathbf{W}^\top\mathbf{A} = \frac{1}{3}\left(\begin{array}{cc|cc|cc} -2 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -2 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 & -2 & 1 \\ 1 & -2 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & -2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -2 \end{array}\right).$$

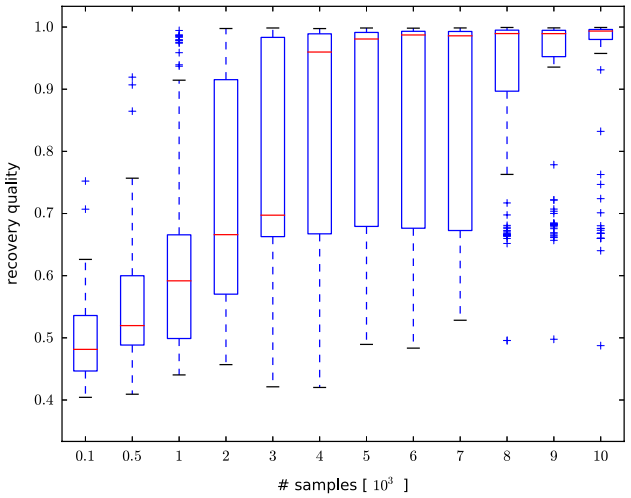


Fig. 2. Recovery quality in dependence of the number of samples. The mixing matrix **A** was uniformly sampled from $O(6)$, after which Algorithm 1 was run on the mixed data **AS**, resulting in a recovery matrix **W**. The recovery quality is given by the block-wise Frobenius-norm of the **WA**. Perfect recovery (a product of at most blockwise permutation and a block diagonal matrix) corresponds to a recovery quality of 1, the theoretically worst recovery index is $1/3$ and the average random recovery is about 0.48. Depicted are the statistics over 100 runs, the boxes extending from lower to upper quartile. For 4000 samples and more, in over 50% of the cases the recovery is virtually perfect.

In this case the recovery quality is $1/3$. Other choices of equally bad $\mathbf{W}^\top\mathbf{A}$ can of course be constructed by permuting blocks of rows (or columns) and by multiplication with a block-diagonal orthogonal matrix.

Fig. 1(c) depicts a typical reconstruction of the recovered subspaces. Statistics of the reconstruction quality over 100 runs of various sample sizes are given in Fig. 2.

5. Discussion and conclusion

The fact that ICA is only applicable to random vectors with a factorization into 1-dimensional components has given rise to the development of generalizations where certain dependencies in the random vectors are allowed. This question was originally raised in [7], where it was argued that ICA algorithms make the data set as independent as possible, after which one simply has to gather the linear components together that turn out still to be dependent after this step, and thus get what then was called Multidimensional Independent Component Analysis (MICA). At this point it was not clear if the nice separability property of ICA holds in this setting, as no formal proof for this claim was available. One of the main contributions of a formal proof for higher dimensional generalizations of ICA was the proof of separability of k-ISA [35], where all subspaces were assumed to have the same size. The proof of this work was based on a multivariate extension of the Darmois–Skitovitch theorem, where the random vectors in question have fixed size k, and the additional assumption is made that the mixing matrix is k-admissible, that is, the aligned $k \times k$ submatrices are either invertible or zero; however this theorem cannot be used in the fully general context where the subspaces are allowed to have arbitrary dimensionality.

Dropping all assumptions on subspace sizes, it is easy to see that any given random vector can be represented by a linear mixture of statistically independent random vectors. The additional assumption of irreducibility of these random vectors restricts them to essentially only one set of such subspaces, so these can be seen as *the* linear factors, or subspaces that together form the original random vector. In any linear mixing model of random vectors it is therefore now possible to assume that the random vector is given as the unique representation of irreducible subspaces, thus easing further analysis.

While this may be an interesting uniqueness result from the pure point of statistics, one may also see this as a structural result: it is now clear that the density of any random vector **X** with existing covariance can be represented uniquely as a product

$$p(\mathbf{x}) = p_1(\mathbf{A}_1\mathbf{x}) \cdots p_n(\mathbf{A}_n\mathbf{x}) p_{\mathrm{Gauss}}(\mathbf{A}_G\mathbf{x})$$

up to the ambiguities discussed (permutation and linear factors). This result has a wide range of applications, e.g. in the fields of signal processing, biomedical imaging and analysis of financial data. Independent Component Analysis (ICA) has proven to be a valuable tool here; see e.g. [9,18,41]. If one assumes that the data set to be analyzed is generated by linear mixings of some independent, unknown underlying sources, which can be modeled by random variables, the fact that ICA is separable tells us that any demixing of the observations into independent components recovers the real sources (up to permutation and scaling). From a theoretical data analysis point of view, the assumption of independence of the sources is very restrictive: even if one knows that the observations are linear mixings of the sources, what happens if the sources are not fully independent? Does the whole idea of ICA completely break down then if one has non-trivial dependencies? As we

have seen, this is not the case, and we actually have a straight-forward extension of ICA to ISA, replacing the independent components with independent irreducible subspaces of arbitrary dimension. It is worthwhile to point out that for random vectors $\mathbf{S}$ fulfilling the ICA assumption of complete mutual independence, one can show that a representation of $\mathbf{S}$ that is merely pairwise independent already is equivalent to $\mathbf{S}$ itself (that is, there is a one-to-one correspondence between the pairwise independent random vectors and the components of $\mathbf{S}$). When showing uniqueness of what is nowadays more commonly referred to as ISA, we assumed mutual independence of the irreducible subspaces, but the open question whether pairwise independence is sufficient is interesting. But even before the question of uniqueness of (mutual) ISA was answered satisfactorily, it already gave rise to the development of algorithms [19,23,27], performing the ISA task. Similar extensions had already been performed for Principal Component Analysis (PCA) – which only employs statistics up to the second order, i.e. correlations – where techniques are known that extract not only a single principal component, but rather a whole principal component subspace [25]. As PCA uses only first and second order moments, it is impossible to uniquely define self-consistent independent subspaces in this context. Therefore, methods to extract such subspaces here have to make use of additional information e.g. by ordering the principal components by power (variance) and then extracting the subspace of the strongest k components. On the other hand, ISA makes use of all higher order statistics, so in this context the term "subspace" actually has a unique meaningful denotation and the sizes of the subspaces arise purely from the definition of independence and irreducibility. The conjecture that ISA can be solved by applying standard ICA algorithms and grouping the recovered random variables that then still show dependencies has been shown for some distributions [31]. Based on this conjecture ISA algorithms performing joint block diagonalization [34,36] or making use of the fact that algorithmic outputs will have a large variability within the subspaces [41] have been developed, but so far there is no algorithmic approach where full convergence in the general setting has been proven.

When considering practical implementations, it is important to point out another fact. In our theoretical work, we have assumed perfect knowledge of the distribution of the given signal, whereas for practical considerations one can only estimate the distribution of the signal up to some precision determined among other things by the number of available samples. In practice one can therefore not expect to estimate full independence of subspaces, even if from a theoretical point of view these actually are independent, e.g. if they represent different independent biophysical processes in the human body. Even in the ICA case, this is an important question and has found attention only recently [12,17,26] although in this setting the original question of course allows a slightly easier handling, as one simply assumes to know the number of subspaces and their dimensionality (in ICA one deals only with 1-dimensional subspaces). For ISA this question has to deserve a lot more attention. Theory answers the question of uniqueness without consideration of the sizes and the number of subspaces, so it would be good to have algorithms that do the same. Furthermore, similar to PCA, for data analysis, the estimation of the dimension of relevant subspaces is non-trivial and we can expect to generalize some of the many existing approaches developed in this easier setting, such as Minimum Description Length (MDL) [29] or Akaike's Information Criterion (AIC) [2] in PCA.

Acknowledgments

The authors acknowledge financial support by the German Ministry for Education and Research (BMBF) via the Bernstein Center for Computational Neuroscience (BCCN) Göttingen under Grant No. 01GQ0430. They thank Florian Blöchl and Claudia Czado for careful proofreading and valuable comments.

Appendix

Proof of Lemma 3.5. Let us first fix two indices i and j in different subspaces. More formally, we fix two different subspace indices $1 \leq i' < j' \leq M$ and then take any i and j such that $\sum_{k=1}^{i'-1} m_k < i \leq \sum_{k=1}^{i'} m_k$ and $\sum_{k=1}^{j'-1} m_k < j \leq \sum_{k=1}^{j'} m_k$. Then, if F is the function defined on the left hand side of Eq. (4),

$$F(\mathbf{x})(\partial_i \partial_j F)(\mathbf{x}) - (\partial_i F)(\mathbf{x})(\partial_j F)(\mathbf{x}) = 0$$

as x_i and x_j appear in different factors of F. So the result of the same operations applied to the right hand side also is 0. Using Lemma 2.1 this tells us that

$$0 = \sum_{k=1}^{M} \left(\prod_{l \neq k} g_k(\mathbf{A}_k \mathbf{x}_k) \right)^2 \left[g_k(\mathbf{A}_k \mathbf{x}) \big(\partial_i \partial_j (g_k \circ \mathbf{A}_k)\big)(\mathbf{x}) - \big(\partial_i (g_k \circ \mathbf{A}_k)\big)(\mathbf{x}) \big(\partial_j (g_k \circ \mathbf{A}_k)\big)(\mathbf{x}) \right].$$

So the right hand side of Eq. (4) is also equal to zero for all i, j fulfilling the inequalities above. Performing this for all such i, j, that is, for all i in the i'-th subspace and for all j in the j'-th subspace, and collecting the expressions into a single matrix and substituting $\mathbf{y}_i := \mathbf{A}_i \mathbf{x}$, we get after using Lemma 2.2

$$0 = \sum_{k=1}^{M} \left(\prod_{l \neq k} g_k(\mathbf{y}_k) \right)^2 \mathbf{A}_{ki'}^{\top} \left[g_k(\mathbf{y}_k) \mathbf{H}_{g_k}|_{\mathbf{y}_k} - (\mathbf{d}g_k|_{\mathbf{y}_k})^{\top} (\mathbf{d}g_k|_{\mathbf{y}_k}) \right] \mathbf{A}_{kj'}$$

for all $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_M)$ that can be written as $\mathbf{y} = \mathbf{Ax}$. Due to the invertibility of $\mathbf{A}$ this is the case everywhere. The functions g_k are twice continuously differentiable, hence continuous. Let us fix $\mathbf{y}_1, \ldots, \mathbf{y}_N$ where for all $k \in \{1, \ldots, M\}$ we have $g_k(\mathbf{y}_k) \neq 0$, which, due to continuity, is then also the case in some environment of the $\mathbf{y}_k$, where we may choose a complex logarithm. Then, with Lemma 2.3, locally

$$\begin{aligned} 0 &= \sum_{k=1}^{M} \left(\prod_{l=1}^{M} g_k(\mathbf{y}_k) \right)^2 \mathbf{A}_{ki'}^{\top} \mathbf{H}_{\log(g_k)}|_{\mathbf{y}_k} \mathbf{A}_{kj'} \\ &= \sum_{k=1}^{M} \mathbf{A}_{ki'}^{\top} \mathbf{H}_{\log(g_k)}|_{\mathbf{y}_k} \mathbf{A}_{kj'} \end{aligned}$$

where the last equality holds as $g_k(\mathbf{y}_k) \neq 0$ for all k. □

References

[1] K. Abed-Meraim, A. Belouchrani, Algorithms for joint block diagonalization, in: Proc. EUSIPCO 2004, January 2004.
[2] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (6) (1974) 716–723.
[3] M. Almeida, R. Vigário, J. Bioucas-Dias, Phase locked matrix factorization, in: Proc. EUSIPCO 2011, June 2011, pp. 1728–1732.
[4] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Computation 7 (6) (1995) 1129–1159.
[5] G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, K.-R. Müller, In search of non-Gaussian components of a high-dimensional distribution, Journal of Machine Learning Research (JMLR) 7 (2006) 247–282.
[6] G. Blanchard, M. Sugiyama, M. Kawanabe, V. Spokoiny, K.-R. Müller, Non-Gaussian component analysis: a semi-parametric framework for linear dimension reduction, in: Proc. NIPS 2005, January 2006, pp. 131–138.
[7] J.-F. Cardoso, Multidimensional independent component analysis, in: Proc. ICASSP, vol. 4, 1998, pp. 1941–1944.
[8] A.Q. Carretero, S.P. Wilson, Dependent gaussian mixture models for source separation, in: Proc. EUSIPCO 2011, June 2011, pp. 1723–1727.
[9] A. Cichocki, S.-I. Amari, Adaptive Blind Signal and Image Processing, John Wiley & Sons, 2002.
[10] P. Comon, Independent component analysis, a new concept? Signal Processing 36 (3) (1994) 287–314.
[11] G. Darmois, Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire, Revue de l'Institut International de Statistique/Review of the International Statistical Institute 21 (1–2) (1953) 2–8.
[12] S.C. Douglas, Z. Yuan, E. Oja, Average convergence behavior of the FastICA algorithm for blind source separation, in: Proc. ICA 2006, in: Lecture Notes in Computer Science, vol. 3889, 2006, pp. 790–798.
[13] J. Eriksson, V. Koivunen, Identifiability and separability of linear ICA models revisited, in: Proc. ICA 2003, 2003, pp. 23–27.
[14] H.W. Gutch, J. Krumsiek, F.J. Theis, An ISA algorithm with unknown group sizes indentifies meaningful clusters in metabolomics data, in: Proc. EUSIPCO 2011, 2011, pp. 1733–1737.
[15] H.W. Gutch, F.J. Theis, Independent subspace analysis is unique, given irreducibility, in: Proc. ICA 2007, in: Lecture Notes in Computer Science, vol. 4666, 2007, pp. 49–56.
[16] W. Härdle, L. Simar, Applied Multivariate Statistical Analysis, Springer Verlag, 2007.
[17] J.M. Herrmann, F.J. Theis, Statistical analysis of sample-size effects in ICA, in: Proc. IDEAL 2007, in: Lecture Notes in Computer Science, vol. 4881, 2007, pp. 416–425.
[18] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
[19] A. Hyvärinen, U. Köster, FastISA: a fast fixed-point algorithm for independent subspace analysis, in: Proc. ESANN 2006, 2006, pp. 371–376.
[20] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Computation 9 (7) (1997) 1483–1492.
[21] A.M. Kagan, Y.V. Linnik, C.R. Rao, Characterization Problems in Mathematical Statistics, Wiley, New York, 1973.
[22] T. Maehara, K. Murota, Algorithm for error-controlled simultaneous block-diagonalization of matrices, SIAM Journal on Matrix Analysis and Applications 32 (2) (2011) 605–620.
[23] Y. Nishimori, S. Akaho, M.D. Plumbley, Riemannian optimization method on the flag manifold for independent subspace analysis, in: Proc. ICA 2006, in: Lecture Notes in Computer Science, vol. 3889, 2006, pp. 295–302.
[24] K. Nordhausen, H. Oja, Scatter matrices with independent block property and ISA, in: Proc. EUSIPCO 2011, June 2011, pp. 1738–1742.
[25] E. Oja, Neural networks, principal components, and subspaces, International Journal of Neural Systems 1 (1) (1989) 61–68.
[26] E. Ollila, H.-J. Kim, V. Koivunen, Compact Cramér–Rao bound expression for independent component analysis, IEEE Transactions on Signal Processing 56 (4) (2008) 1421–1428.
[27] B. Póczos, A. Lőrincz, Independent subspace analysis using k-nearest neighborhood distances, in: Proc. ICANN 2005, in: Lecture Notes in Computer Science, vol. 3697, 2005, pp. 163–168.
[28] B. Póczos, Z. Szabó, J. Schneider, Nonparametric divergence estimators for independent subspace analysis in: Proc. EUSIPCO 2011, 2011, pp. 1718–1722.
[29] J. Rissanen, Modeling by shortest data description, Automatica 14 (5) (1978) 465–471.
[30] V.P. Skitovich, Linear forms of independent random variables and the normal distribution law, Izvestiya Akademii Nauk SSSR. Seriya Matematicheskaya 18 (2) (1954) 185–200.
[31] Z. Szabó, B. Póczos, A. Lőrincz, Separation theorem for $\mathbb{K}$-independent subspace analysis with sufficient conditions, January 2006. arXiv:math/0608100 [math.ST].
[32] R.E. Tarjan, Depth-first search and linear graph algorithms, in: 12th Annual Symposium on Switching and Automata Theory 1971, 1971, pp. 114–121.
[33] F.J. Theis, A new concept for separability problems in blind source separation, Neural Computation 16 (9) (2004) 1827–1850.
[34] F.J. Theis, Blind signal separation into groups of dependent signals using joint block diagonalization, in: Proc. ISCAS 2005, Jaunary 2005, pp. 5878–5881.
[35] F.J. Theis, Multidimensional independent component analysis using characteristic functions, in: Proc. EUSIPCO 2005, Jaunary 2005.
[36] F.J. Theis, Towards a general independent subspace analysis, in: Proc. NIPS 2006, Jaunary 2007, pp. 1361–1368.
[37] F.J. Theis, A. Jung, C.G. Puntonet, E.W. Lang, Linear geometric ICA: fundamentals and algorithms, Neural Computation 15 (2) (2003) 419–439.
[38] F.J. Theis, M. Kawanabe, Uniqueness of non-Gaussian subspace analysis, in: Proc. ICA 2006, in: Lecture Notes in Computer Science, vol. 3889, 2006, pp. 917–925.
[39] A. Yeredor, Blind source separation via the second characteristic function, Signal Processing 80 (5) (2000) 897–902.
[40] J. Ylipaavalniemi, R. Vigário, Subspaces of spatially varying independent components in fMRI, in: Proc. ICA 2007, in: Lecture Notes in Computer Science, vol. 4666, 2007, pp. 665–672.
[41] J. Ylipaavalniemi, R. Vigário, Analyzing consistency of independent components: an fMRI illustration, NeuroImage 39 (1) (2008) 169–180.